Hazel Choe

Professor Bilen

Data 400: Data Analytics Capstone

11 September 2025

Project Proposal: Do Google Search Trends Predict Best-Selling Albums?

The rise in search activity has greatly contributed to consumer interest before sales results are released. However, we do not know if the sales results are actually connected with online attention, such as search activity. The objective of this project is to determine whether stronger or earlier search activity increases album sales by applying machine learning algorithms and logistic regression to a combined dataset. In the end, it would assist stakeholders in forecasting album performance and allocating marketing budgets. However, artists may gain a deeper understanding of the relationship between online attention and sales results, and cultural researchers can better comprehend how online attention influences long-term music trends.

The first step will be scraping the best-selling albums of all time data from chartmasters.org. Next, gather the Top 100 artists and albums from the scraped data, and search activity data on those artists and albums. Search activity data will be collected from Google Trends. To ensure consistency between sources, the two datasets will be merged by matching album titles and artist names. The time range will be adjusted based on whether search spikes occurred before, during, or after album chart entries to capture the relationship between consumer interest and sales volume. This process will generate a combined dataset integrating both online search activity and commercial performance.

Logistic Regression is suitable for this analysis as it allows for binary classification and produces interpretable coefficients indicating whether stronger or earlier search activity significantly increases the likelihood of an album entering the charts. The independent variables are the duration of sustained interest, the timing of the search surge, and the Google search peak value. The dependent variable is whether the album enters the best-selling album Top 100 list. For example, according to the logistic regression analysis, albums with a search interest level of 70 out of 100 or higher for at least two weeks before release have a significantly higher probability of entering the best-selling album Top 100.

Addressing ethical and social considerations is crucial. All data will be collected from publicly accessible, legitimate sources, and the use of Google Trends will comply with its terms of service. It is important to note that search data may not fully represent all demographics, as internet access and search patterns vary by region and age group. Furthermore, this analysis will only examine the correlation between sales volume and consumer interest; it will not address the relationship between popularity and artistic value.

Finally, this project is engaging and original by linking two related indicators of popularity: album sales and internet search activity. Rather than simply listing the best-selling albums, it asks whether public interest can predict commercial success. Data collection and scraping are manageable, the methodology is suitable for a small-scale project, and the insights are relevant both practically and academically.